RF-Net 2: Fast Inference of Virus Reassortment and Hybridization Networks – Supplemental Material

Alexey Markin
Virus and Prion Research Unit
National Animal Disease Center, USDA-ARS, USA

Sanket Wagle
Department of Computer Science,
Iowa State University, USA

Tavis K. Anderson
Virus and Prion Research Unit
National Animal Disease Center, USDA-ARS, USA

Oliver Eulenstein
Department of Computer Science,
Iowa State University, USA

S1 Simulation of alignments

To obtain realistic GTR+Gamma model parameters we downloaded publicly available (complete) HA sequences of the 1A classical swine lineage collected starting 2015 from Influenza Research Database (`fludb.org` accessed on 10/13/2021) (Zhang *et al.*, 2017). We aligned all sequences ($n = 2248$) with MAFFT v7.475 (Katoh and Standley, 2013) and trimmed them to the coding region. We then estimated a maximum likelihood tree with IQ-Tree v.1.6.12 (Nguyen *et al.*, 2015). To sample the variation in substitution rates across different parts of the tree, we isolated all minimal subtrees from the global phylogeny with at least 50 taxa. This procedure resulted in 19 independent subtrees. For each subtree we estimated relative transition rates for the GTR model, nucleotide base frequencies, and the gamma-shape parameter alpha using RAxML v.8.2.12 (Stamatakis, 2014). We then fitted the observed 19 vectors of relative transition rates to a Dirichlet distribution using maximum likelihood, which resulted in distribution $Dirichlet(4.448, 18.700, 1.034, 0.757, 20.660, 2.603)$. Further, we fitted the observed base frequency vectors to a Dirichlet distribution, which resulted in $Dirichlet(4317.595, 2308.357, 2732.760, 3007.051)$. Finally, we fitted the observed alpha parametrs to a log-normal distribution, resulting in $Lognormal(-1.412, 0.955)$.

Then, to simulate sequence alignments we sampled the relative transition rates, base frequencies, and gamma shapes from the above distributions independently for each simulated gene.

S2 Supplemental figures

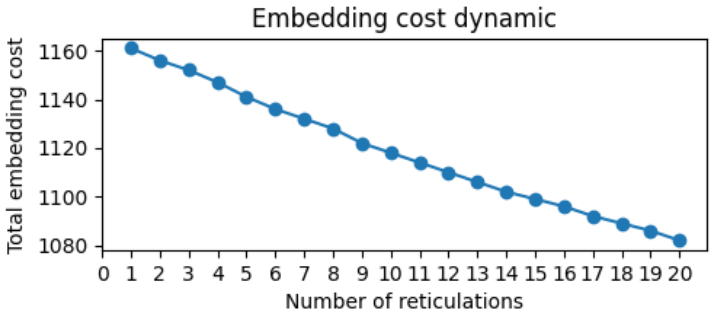

Figure S1: Dynamic decrease in embedding cost associated with the number of reticulations in H3.2010.1 reticulation network.

References

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, **30**(4), 772–780.

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, **32**(1), 268–274.

Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**(9), 1312–1313.

Zhang, Y., Aevermann, B., Anderson, T., Burke, D., Dauphin, G., Gu, Z., He, S., Kumar, S., Larsen, C., Lee, A., *et al.* (2017). Influenza research database: An integrated bioinformatics resource for influenza virus research. *Nucleic acids research*, **45**(D1), D466.